DOCUMENT RESUME

ED 395 015 TM 025 029

AUTHOR	Mislevy, Robert J.
TITLE	Foundations of a New Test Theory.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Office of Naval Research, Arlington, VA. Cognitive and Neural Sciences Div.
REPORT NO	ETS-RR-89-52-ONR
PUB DATE	Oct 89
CONTRACT	N00014-88-K-0304
NOTE	37p.
PUB TYPE	Viewpoints (Opinion/Position Papers, Essays, etc.) (120)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Ability; Cognitive Psychology; *Educational Assessment; Educational Research; Elementary Secondary Education; Estimation (Mathematics); Higher Education; Instruction; *Problem Solving; *Psychological Studies; *Statistical Analysis; Student Placement; *Test Theory

ABSTRACT

It is only a slight exaggeration to describe the test theory that dominates educational measurement today as the application of twentieth-century statistics to nineteenth-century psychology. Sophisticated estimation procedures, new techniques for missing-data problems, and theoretical advances into latent-variable modeling have appeared--all applied with psychological models that explain problem-solving ability in terms of a single, continuous variable. This caricature suffices for many practical prediction and selection problems because it expresses patterns in data that are pertinent to the decisions that must be made. It falls short for placement and instruction problems based on students' internal representations of systems, problem-solving strategies, or reconfigurations of knowledge as they learn. Such applications demand different caricatures of ability--more realistic ones that can express patterns suggested by recent developments in cognitive and educational psychology. The application of modern statistical methods with modern psychological models constitutes the foundation of a new test theory. (Contains 99 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-89-52-ONR

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED 395 015

FOUNDATIONS OF A NEW TEST THEORY

Robert J. Mislevy

This research was sponsored in part by the
Cognitive Science Program
Cognitive and Neural Sciences Division
Office of Naval Research, under
Contract No. N00014-88-K-0304
R&T 4421552

Robert J. Mislevy, Principal Investigator

Educational Testing Service
Princeton, New Jersey

October 1989

Reproduction in whole or in part is permitted
for any purpose of the United States Government.

Approved for public release; distribution unlimited.

TM025029

BEST COPY AVAILABLE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION: Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S): RR-89-52-ONR	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION: Educational Testing Service; 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION: Cognitive Science Program, Office of Naval Research (Code 1142CS), 800 North Quincy Street
6c. ADDRESS (City, State, and ZIP Code): Princeton, NJ 08541	7b. ADDRESS (City, State, and ZIP Code): Arlington, VA 22217-5000
8a. NAME OF FUNDING/SPONSORING ORGANIZATION; 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: N00014-88-K-0304
8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
61153N	RR04204	RR04204-01	R&T4421552

11. TITLE (Include Security Classification): Foundations of a New Test Theory (Unclassified)

12. PERSONAL AUTHOR(S): Robert J. Mislevy

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical	FROM _____ TO _____	October 1989	34

16. SUPPLEMENTARY NOTATION

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP
05	10	

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number): Cognitive psychology; Psychometrics; Educational measurement; Test theory; Item response theory

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

It is only a slight exaggeration to describe the test theory that dominates educational measurement today as the application of twentieth century statistics to nineteenth century psychology. Sophisticated estimation procedures, new techniques for missing-data problems, and theoretical advances into latent-variable modeling have appeared--all applied with psychological models that explain problem-solving ability in terms of a single, continuous variable. This caricature suffices for many practical prediction and selection problems because it expresses patterns in data that are pertinent to the decisions that must be made. It falls short for placement and instruction problems based on students' internal representations of systems, problem-solving strategies, or reconfigurations of knowledge as they learn. Such applications demand different caricatures of ability--more realistic ones that can express patterns suggested by recent developments in cognitive and educational psychology. The application of modern statistical methods with modern psychological models constitutes the foundation of a new test theory.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION	
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	Unclassified	
22a. NAME OF RESPONSIBLE INDIVIDUAL: Dr. Charles E. Davis	22b. TELEPHONE (Include Area Code): 202-696-4046	22c. OFFICE SYMBOL: ONR 1142CS

DD Form 1473, JUN 86 — Previous editions are obsolete. S/N 0102-LF-014-6603

SECURITY CLASSIFICATION OF THIS PAGE: Unclassified

Foundations of a New Test Theory

Robert J. Mislevy

Educational Testing Service

October 1989

To appear in N. Frederiksen, R. Mislevy, and I. Bejar (Eds.), Test Theory for a New Generation of Tests, Lawrence Erlbaum and Associates. This work was supported in part by Contract No. N00014-88-K-0304, R&T 4421552, from the Cognitive Science Program, Cognitive and Neural Sciences Division, Office of Naval Research. I am grateful to Henry Braun, Norman Frederiksen, and Kentaro Yamamoto for comments on earlier drafts.

Foundations of a New Test Theory

Abstract

It is only a slight exaggeration to describe the test theory that dominates educational measurement today as the application of twentieth century statistics to nineteenth century psychology. Sophisticated estimation procedures, new techniques for missing-data problems, and theoretical advances into latent-variable modeling have appeared--all applied with psychological models that explain problem-solving ability in terms of a single, continuous variable. This caricature suffices for many practical prediction and selection problems because it expresses patterns in data that are pertinent to the decisions that must be made. It falls short for placement and instruction problems based on students' internal representations of systems, problem-solving strategies, or reconfigurations of knowledge as they learn. Such applications demand <u>different</u> caricatures of ability--more realistic ones that can express patterns suggested by recent developments in cognitive and educational psychology. The application of modern statistical methods with modern psychological models constitutes the foundation of a new test theory.

Key Words: Cognitive psychology
Educational measurement
Item response theory
Psychometrics
Test theory

Introduction

Educational measurement faces a crisis today that would appear to threaten its very foundations. The essential problem is that the view of human abilities implicit in standard test theory--item response theory as well as classical true-score theory--is incompatible with the view rapidly emerging from cognitive and educational psychology. Learners increase their competence not by simply accumulating new facts and skills, but by reconfiguring their knowledge structures, by automating procedures and chunking information to reduce memory loads, and by developing strategies and models that tell them when and how facts and skills are relevant. The types of observations and the patterns in data that reflect the ways that students think, perform, and learn cannot be accommodated by traditional models and methods. To some it would seem to some that psychometrics has little to offer in the quest to apply this new knowledge to the practical educational problems of the individual, the classroom, or the nation (Hunt and MacLeod, 1978).

I concur that the standard methods of test theory do not suffice for solving problems cast in the framework of what we are learning about how people acquire knowledge and competence, but I cannot agree that psychometrics has nothing to offer.

Standard test theory evolved as the application of statistical theory with a simple model of ability that suits the decision-making environment of most mass educational systems. Broader educational options, based on insights into the nature of learning and supported by more powerful technologies, demand a broader range of models of capabilities--still simple compared to the realities of cognition, but capturing patterns that inform a broader range of alternatives. A new test theory can be brought about by applying to well-chosen cognitive models the same general principles of statistical inference that led to standard test theory when applied to the simple model.

The first half of this paper sketches the evolution of standard test theory, highlighting the challenges that spurred each new advance. The challenges that cognitive and educational psychology present today are then discussed, and a framework for responding to that challenge is outlined. Directions for needed development are exemplified with current work.

The Early Context of Educational Decisions

The kinds of decisions that shaped the evolution of classical test theory were nearly universal in education at the beginning of this century, and dominate practice yet today. They were born of the constraints educators encountered as they launched their campaign to provide education on a broader scale than had ever been attempted hitherto:

> "...the demand for tests arose during the period when school attendance was made compulsory and when higher education was developing its strengths. Educators faced the unprecedented dilemma of dealing with the range and diversity of abilities and backgrounds that individuals bring to schooling. They needed ways of determining which children and youths would be able to profit from some form of instruction as given in ordinary school and college practices as designed essentially for the majority of the population." (Glaser, 1981, p. 924).

Educators were confronted with selection or placement decisions for large numbers of students. Resources limited the information they could gather about each student, constrained the number of options they could offer, and precluded tailoring programs to individual students once a decision was made.

A first example is selecting applicants into a college that presents the same material in the same way to all students. There is only one treatment, and the alternatives are to accept or reject. The admissions officer would prefer to accept those who are likely to succeed. When resources permit more than one decision option, the usual generalization of the accept/reject paradigm is to offer a sequence of alternatives, each more demanding than the next. Placing high school freshmen into

academic tracks is an example of this latter type. Problems of selection into a single program and of placement into a single sequence are both decisions about "linearly ordered options."

Exposing a diverse group of students to a uniform educational treatment typically produces a distribution of outcomes (Bloom, 1976). An individual's degree of success depends on how his or her unique skills, knowledge, and interests match up with the equally multifaceted requirements of the treatment.

At costs substantially lower than personal interviews or performance samples, responses to multiple-choice test items provide information about certain aspects of this matchup. What is necessary is that each item tap some of the skills required for success. Even though a single item might require only a few of the relevant skills and offer little information in its own right, a tendency to provide correct answers over a large number of items supports some degree of prediction of success (Green, 1978). If all candidates are administered the same items, and one wishes to predict success in linearly-ordered options, their number-correct scores can be used (Dawes and Corrigan, 1974). Even though the several students at a given score level possess different constellations of skills, abilities, and backgrounds, making the same decision for all of them <u>among the available alternatives</u> is often about as well as can be done <u>with the available data</u>.

Once the test and the linearly-ordered options are specified, making decisions from test performances requires nothing more complicated than adding up numbers of correct responses. Two different tests constructed for the same decision, however, invariably line up examinees differently as they draw upon different particular skills from the myriad of those potentially informative. Additional statistical machinery is required to guide one in constructing tests and evaluating their quality. Classical test theory was a first response to these needs.

Classical Test Theory

Charles Spearman (1904a, 1904b, 1907, 1910, 1913) is credited with the central idea of classical test theory (CTT): a test score can be viewed as the sum of two components, a "true" score and a random "error" term. Two similar ("parallel") tests are considered to reflect the same true score, but disagree about an examinee's observed scores because of the error components--the variance of which can, under the assumptions of CTT, be driven to zero by just making the tests long enough. Ideally decisions would be based on true scores; in practice they must be based on observed scores. "Reliability," the degree to which the unobservable true scores account for the variance in observed scores, gauges the accuracy with which a test lines up a group of examinees--a reasonable criterion for the quality of a test if it is assumed that the items tap appropriate skills and scores will be used to decide among linearly ordered options.

Upon these notions was founded a practicable testing methodology. Reliability became a paramount measure of the quality of a test, although of course reliability had to be complemented with validity measures such as the correlation between test scores and subsequent performance. Validity studies had less influence on test construction, however, because they arrive too late in the process--only after the test has been administered and examinees have been followed over time. To obtain high reliability, one uses items that would be answered correctly by about half the examinees, for example, and avoids items that would have low correlations with the total test scores.

Note that these dicta could guide test construction solely from counts and patterns of right and wrong responses to candidate test items--ignoring both the content of the items and the contemplated decision alternatives. Of course good test construction does consider the knowledge, skill, and strategy requirements of items. The point is that these considerations lie outside the realm of the classical test theory. Test developers

use them independently of, sometimes in contradiction to, what test theory tells them.

Building upon Spearman's foundation, psychometricians developed a vast armamentarium of techniques for building and using tests (Gulliksen, 1950), such as approximating reliability from the internal consistency of items within a test (Kuder and Richardson, 1937) and estimating validity without knowing subsequent performances of rejected examinees (Kelley, 1923). Over time, a rigorous axiomatic foundation was laid for statistical inference under the aegis of CTT (Lord, 1959; Novick, 1966; Lord and Novick, 1968). The simple partitioning of observed scores into true and error components was generalized to multiple sources of variation from items, persons, and observational settings, and the full power of analysis of variance was brought to bear upon decision-making problems using test scores (Cronbach, Gleser, Nanda, and Rajaratnam, 1972; Lord and Novick, 1968).

A source of dissatisfaction with CTT early on was that its characterizations of examinees, such as total score and percentile rank, and of items, such as percent-correct and item-test correlation, are confounded descriptions of the particular items that constitute a test and a particular group of examinees who takes it (Wright, 1968). If one test consists of easier items than a second otherwise similar test, examinees' scores on the two tests are not directly comparable and score distributions have different shapes. If a test is administered to groups of examinees that differ in proficiency, item percents-correct and item-test correlations differ. When many tests could be constructed for the same purpose, differing perhaps in difficulty or length, should not there be a way to characterize examinees independently of the test they took, and items independently of the examinees who took them?

In attitude measurement, where agreements to a topic are analogous to correct answers to test questions, L.L. Thurstone (1928) expressed the following desideratum: "If a scale is to be regarded as valid, the scale values of the statements should not

be affected by the opinions of the people [whose responses] help to construct it." Thurstone (1925) and E.L. Thorndike (Thorndike et al., 1926) pioneered efforts to relate test scores to psychological traits, using item percents-correct and assumptions about distributions of traits to transform scores from different tests onto the same scale.

Thurstone and Thorndike scaling, despite allusions to an underlying trait, remained essentially theories for scores, albeit transformed (with the aid of untestable assumptions) to permit comparisons across nonparallel tests. Psychological traits per se appear as explicit parameters in the models of Ferguson (1942), Lawley (1943), and Tucker (1946). These researchers studied test construction problems within CTT by making an assumption beyond those of CTT proper; namely, that aside from random factors, item responses were driven by a unobservable ability variable. A second generation of test theory began to take form as attention shifted from test scores as the object of inference, to unobservable variables hypothesized to have produced them.

Item Response Theory

Item response theory (IRT), or, "latent trait theory," as it was called then, appears as a test theory in its own right in the work of Frederic Lord (1952) and Georg Rasch (1960). Like classical test theory, IRT concerns examinees' overall proficiency in a domain of tasks. But while CTT makes no statement about the mechanisms that give rise to performance, IRT posits a single, unobservable, proficiency variable.[1]

At the heart of IRT is a mathematical model for the probability that a given person will respond correctly to a given

---

[1] If classical test theory offers a statistical model for test scores without a psychological model, Guttman's (1944) scaling techniques offer a psychological model without a statistical model. Important in the reconceptualization of the meaning of test scores, a Guttman scale can be viewed as the limiting case in IRT in which each item is perfectly informative about whether an examinee's ability lies above or below a specific point on an ability continuum.

item, a function of that person's proficiency parameter and one or more parameters for the item. The item's parameters express properties such as difficulty or sensitivity to proficiency. The item response, rather than the test score, is the fundamental unit of observation. If an IRT model holds, responses to any subset of items support inferences on the same scale of measurement.

This conceptualization opens the door to solving many practical testing problems that were difficult under CTT, such as:

Test construction (Birnbaum, 1968; Theunissen, 1985). If item parameters are available for a collection of items, tests can be constructed for optimal performance in specific applications, such as minimizing classification errors.

Adaptive testing (Lord, 1980, Chapter 10; Weiss, 1984). An adaptive testing scheme selects the best item to administer next to an examinee, based on the amount of information that various available items would provide and a provisional estimate of the examinee's proficiency from responses to items given thus far.

Educational assessment (Bock, Mislevy, and Woodson, 1982; Choppin, 1976; Messick, Beaton, and Lord, 1983). Assessments gauge proficiencies at the level of populations rather than individuals, to evaluate programs and monitor trends. IRT makes it possible to establish a stable measurement scale while allowing assessment instruments to evolve over time.

This work assumed, for the most part, that the IRT model was known and correct, and that true values or accurate estimates of item parameters were available. Current IRT research emphasizes integrating IRT into the general framework of statistical inference, and acquiring an understanding of just when and how IRT models are appropriate.

Statistical Inference in Item Response Theory

Early applications of IRT were designed more to demonstrate its potential than to solve actual measurement problems. Data were gathered with tests written according to CTT dicta; the same long tests were administered to many examinees, and each item had

passed CTT quality checks. Illustrative purposes were served adequately by rough estimation procedures that treat point estimates of examinee- and item-parameters as if they were the parameters themselves, ignoring the uncertainty associated with the estimates. These approximations break down when IRT is applied beyond the usual limits of CTT testing, as when examinees are presented only, say, fifteen items in adaptive testing or five in educational assessments (Mislevy, 1988). In response, IRT researchers have turned to two active lines of research in statistics: missing data methods and Bayesian estimation.

Missing data methods are relevant because a latent variable such as an IRT examinee proficiency parameter can be viewed as a datum whose value is missing for everyone. General results on estimating parameters when some data are missing, such as Dempster, Laird, and Rubin's (1977) EM algorithm, have led to methods of item parameter estimation that are at once rigorous and efficient (e.g., Bock and Aitkin, 1980; Tsutakawa, 1984). Results on statistical information in missing data problems yield insights into the uncertainty structures of IRT parameters (Mislevy and Sheehan, in press; Mislevy and Wu, 1988) and offer ways of increasing accuracy by exploiting collateral information about items and examinees (Mislevy, 1987, 1988a).

The Bayesian perspective confronts uncertainty head on, expressing what is known about parameters as probability distributions. When these distributions are concentrated, the expedient of using point estimates as if they were the true parameters can give acceptable results in subsequent analyses. But when the distributions are diffuse, one must propagate the uncertainty into subsequent analyses to obtain correct inferences. Statistical reasoning along these lines was proposed as far back as 1927 by Kelley (1927), and championed by Novick in the 1970's (e.g., Novick and Jackson, 1974), but only now are the ideas gaining currency. In this framework, one can determine when the standard, simpler, approximations suffice, but use (admittedly more complex) correct analyses when they don't. For examples in

IRT estimation problems, see Bock and Aitkin (1981) on item parameters, Mislevy (1988b) on proficiency distributions, and Tsutakawa and Soltys (1988) on individuals' proficiencies.

The Question of Model Fit

But of course the IRT model is never exactly correct. A single variable that accounts for all nonrandomness in examinees' responses is not a serious representation of cognition, but a caricature that can solve applied problems when it captures the patterns that are salient to the job. The pattern that CTT and IRT can capture is examinees' tendencies to give correct responses, which can usefully inform decisions about linearly ordered alternatives. IRT was a practical advance beyond CTT because it provides information about overall proficiencies in more flexible ways. It was a conceptual advance because it provides a framework for detecting anomalies in the "overall proficiency" paradigm.

This can be illustrated with Rasch's (1960) model for right/wrong items, supposing for convenience all examinees are presented the same test. Under CTT, all examinees with a given total score would be treated alike. Under the Rasch model, all examinees with the same score would receive the same ability estimate[2], and <u>might</u> also be treated alike--depending on an analysis of model fit. Combining an examinee's proficiency estimate with an item's difficulty estimate, the Rasch model states how likely a correct response would be if the single-proficiency conception of ability were true. The items that high scorers missed should usually be easy ones, and the items low scorers got right should be easy ones. Finding that these

---

[2] Under other IRT models such as the 2- and 3-parameter logistic models, examinees with the same total score need not receive <u>exactly</u> the same ability estimate, but usually receive similar estimates. Correlations between total scores and IRT estimates in typical educational tests are usually above .95, and few decisions would be made differently with any IRT model, or, if everyone has taken the same test, even with CTT.

patterns hold supports making the same decisions about people with same scores, because, to an approximation, they got the same items right and the same ones wrong. Total scores, and thus Rasch ability estimates, convey nearly everything these data have to say about comparing these examinees.

To the extent that high scoring examinees miss items that are generally easy and low scoring examinees get hard ones right, neither total scores nor IRT ability estimates may be capturing all the systematic information in the data. Analyses of an individual's unexpected responses can reveal misconceptions or atypical patterns of learning (Mead, 1976; Smith, 1986; Tatsuoka, 1983). To understand these patterns one must look beyond the simple universe of the IRT model--to the content of the items, the structure of the learning area, the pedagogy of the discipline, and the psychology of the problem solving tasks the items demand.

Now, patterns in responses other than overall level proficiency can have educational and psychological meaning, but yet hold no salience for a particular decision. If overall proficiency in a domain of items suffices for a particular decision, as can be the case with linearly ordered educational options, cross-current patterns constitute data variation that need not be explicated. This is the essence of statistical modeling: expressing the patterns that are dominant and meaningful in terms of model parameters, and allowing for departures from these patterns in terms of distributions of residuals. But if the decision does depend on the cross-current patterns, in addition to or instead of overall proficiency, neither CTT nor standard IRT may be the right tool for the job.

The issue of model fit, then, is more pragmatic than statistical, since lack of fit must be judged in practice by the nature and the magnitude of the errors it causes. An IRT model might be satisfactory for selecting honors math students, for example, if people with similar scores have similar chances of success--even though examinees with similar scores have different profiles of skills and knowledge. The profile differences could

be modeled as "noise" without harm for the selection decision--but probably not for advising individual examinees which topics to study to maximally increase their scores.

Measuring learning is one application where IRT models can fail, because they accommodate only a highly constrained type of change: an examinee's chances of success on all items must increase or decrease by exactly the same amount (in an appropriate metric). A single IRT model applied to pretest and posttest data cannot reveal how different students learn different topics to different degrees--patterns that could be at the crux of an instructional decision.

Testing and Learning

Good "macro-level" decisions to place students into appropriate educational programs are important in increasing the quality of education, but they are not sufficient. Tracking students as they progress opens the door to finer grained "micro-level" decisions to enhance learning along the way. Good decision-making at this level requires an inferential framework built around an understanding of how students learn.

A picture of a learner that is consistent with standard test theory is that of a collector of facts and skills, adding each to his repertoire more or less independently of others. Recent developments in psychology sketch a markedly different picture, reflecting the astounding capabilities and the surprising limitations of the mind--lightning fast recognition of stored patterns and creative applications of heuristic strategies, on the one hand; yet with short term memory capacities of only about seven elements and an inability to perform more than one attention-demanding task at a time. Performance is to be understood through the availability of well-practiced procedures that no longer demand high levels of attention ("automaticity"); strategies by which actions are selected, monitored, and, when necessary, switched ("metacognitive skills"); and the mental structures that relate facts and skills ("schema"). Learning is

to be understood through the automatization of procedures; the acquisition and enhancement of metacognitive skills; and the construction, revision, and replacement of schema.

Comparing the performances of novices and experts offers insights into the nature of performance and learning. A first, unsurprising, difference is that experts command more facts and concepts than novices, and have richer interconnections among them. Interconnections overcome limitations of short term memory; while the novice may work with seven distinct elements, the expert works with seven constellations that embody relationships among many elements ("chunking"). Moreover, experts often organize their knowledge in schemata possessing not simply more connections, but qualitatively different ones. The advanced concepts that college physics students acquire, for example, can be organized around informal associations or naive misconceptions (Caramazza, McCloskey, and Green, 1981). These novices tackle physics problems in less effective ways than expert physicists, whose more appropriate schemata lead them to the crux of the matter (Chi, Feltovich, and Glaser, 1981). Experts also differ from novices by having automatized, through study and practice, procedures that were once slow and attention consuming, allowing them to focus on novel aspects of a problem, look from different perspectives, and more efficiently monitor and guide their efforts as they work (Lesgold and Perfetti, 1978).

The challenge to education is to discover what experiences help a learner with a given configuration of propositions, skills, and connections to reconfigure that knowledge into a more powerful arrangement. Vosniadou and Brewer (1987) point to Socratic dialogue and analogy as mechanisms that facilitate such learning. To apply them effectively, one must take into account not simply target configurations, such as the expert's model, but the individual learners' current configurations. The challenge to test theory is to provide models and methods to assess knowledge, and to guide instruction, as seen in this new light.

To what extent can standard test theory meet this challenge? Recall that standard test theory characterizes performance only as to overall level of proficiency, and learning only as to change in overall proficiency. Cronbach and Furby (1970) note the inadequacy of such measures of change when applied with conventional broad range educational tests:

> Even when [test scores] X and Y are determined by the same operation [e.g., scores under the same CTT or IRT model], they often do not represent the same psychological processes (Lord, 1958). At different stages of practice or development different processes contribute to performance of a task. Nor is this merely a matter of increased complexity; some processes drop out, some remain but contribute nothing to individual differences within an age group, some are replaced by qualitatively different processes. (p. 76).

Standard test scores can be connected more closely with cognition if they summarize performance over only tasks that are very homogeneous in their requirements (Glaser, 1963), and this specificity marked the criterion referenced testing movement of the 1960's and 1970's. Merely defining testing areas very narrowly, however, is not sufficient to make test scores instructionally relevant (Glaser, 1981). A list of scores in narrowly defined areas ignores the interconnections among scores induced by the knowledge, skills, and strategies they tap in pairs, in triples, or in hierarchies of the specific behaviors--yet it is at just this level that instructional relevance must be sought.

New Tests, New Test Theory

A learner's state of competence at a given point in time is a complex constellation of facts and concepts, and the networks that interconnect them; of automatized procedures and conscious heuristics, and their relationships to knowledge patterns that signal their relevance; of perspectives and strategies, and the management capabilities by which the learner focuses his efforts.

There is no hope of providing a complete description of such a state. Neither is there a need to. The new pedagogy need merely(!) identify communalities among states of competence that can be linked to instructional actions that facilitate changes to preferable states. Distinctions need not be made among all possible states, but only among classes of states with different instructional implications. The new tests to inform instructional decisions need merely(!) present tasks that learners in the different states are likely to carry out in observably different ways. Not only correctly as opposed to incorrectly, but at what speed, with what intermediate products, or with which incorrect response; not simply as independent pieces of information from distinct items, but in patterns of similarity, dissimilarity, or independence across tasks that probe knowledge structures and problem-solving strategies. The new test theory need merely(!) provide models whose parameters are capable of expressing the salient patterns, and inferential procedures upon which to base instructional decisions in the presence of uncertainty.

Foundations of the new pedagogy are to be found in the union of analyses of key concepts in a substantive area, research into the cognitive psychology of the area, and detailed observations of learners as they progress. Greeno (1976) argues that the tools and the perspectives of cognitive and educational psychology have developed to a point at which they can be used to generate instructional objectives in this manner. He provides detailed illustrations in three substantive domains at increasing levels of complexity and sophistication: fourth-grade fractions, high school geometry, and college level auditory psychophysics.

Foundations of the new theory of test construction are similarly to be found in educational and cognitive psychology (Embretson, 1985a; Messick, 1984). Standard vocabulary items suffice to ascertain the breadth of a learner's familiarity with concepts in a substantive area, but tasks based on analogies probe the interconnections among concepts. Speed of response is more informative than correctness about the automaticity of procedures,

hence a better guide to assigning additional practice on a currently conscious process. Designing appropriate measures demands familiarity with the substantive field, not just about the knowledge structures of the expert but about the incomplete or inaccurate structures novices often use. To see how the requisite cognitive and substantive analyses might be carried out, and how tasks that differentiate among learners at different states of competence might then be constructed, the reader is referred to Curtis and Glaser (1983) on reading achievement and Marshall (1985) on "story problems" in arithmetic.

Foundations of the new test theory are to be found in the general principles that led to the development of item response theory. The examinee will be characterized by parameters that express tendencies to act in accordance with the various continuous levels or discrete states in simplified models of cognition. Tasks will be characterized by parameters indicate the extent to which they tap different aspects of knowledge structures, procedures, or strategies. As in IRT, individual differences among examinees that are not salient to the decision will be modeled as random--not as a psychologically tenable position, but as a practically useful expedient.

Beyond "Low-to-High Proficiency"

The breadth of problems to which standard test theoretic models have been usefully employed, despite their limited low-to-high conception of proficiency, suggests a certain robustness of modeling. It is not necessary that models account for all possible ways students might approach a test, but it is necessary that they can capture instructionally relevant patterns. A test must be designed to highlight the pertinent patterns, and analyzed with a model capable of expressing them.

The idea of building test items around cognitive principles can be traced back at least as far as to Guttman's facet design tests (Guttman, 1970). Guttman worked out analytic methods for analyzing data from such tests within the framework of classical

BEST COPY AVAILABLE

test theory. Scheiblechner (1972) and Fischer (1973), with their "linear logistic test model" expressed item difficulty parameters in the Rasch IRT model as functions of psychologically salient features of test items, but still characterized examinees in terms of overall proficiency. More recently, test theory models built around patterns other than overall proficiency have begun to appear in the psychometric literature.

"Tectonic plate" models. Increasing competence in a substantive area need not be reflected as uniformly increasing chances of success on all tasks. Patterns of smooth increase may be observed for certain people on certain sets of tasks, in certain phases of development; standard test theory will give good summaries of change in these neighborhoods. Discontinuous patterns of change begin to appear as the scope of tasks becomes broader, as the range of development becomes greater, and as the range of experiences of examinees becomes more diverse. "Tectonic plate" models generalize IRT by allowing for a limited number of predetermined, theory-driven, discontinuities in item response patterns. In tectonic plate geological models, points within a given land mass, or plate, maintain their relative positions, but the plates move with respect to one another. In tectonic plate psychometric models, items tapping the same set of skills maintain their difficulties relative to one another, but the difficulties of the groups of items change with respect to other groups as learners acquire new skills or concepts.

Wilson's (1985, 1989) "Saltus" model extends the Rasch IRT model to development with discontinuous jumps. An example is Siegler's (1981) rule-learning analysis of balance-beam tasks, where students can increase their competence either by using the rules they know more effectively (continuous change) or by learning new rules (discontinuous change). Sometimes students who learn a new rule begin to miss a type of problem they used to get right, because their previous, less complete, set of rules gave the right answers for the wrong reasons. This pattern flouts standard test theory. The Saltus model assumes that each examinee

is in one of a number of unobservable stages of development. Items are classified so that all items in a class have the same relationship to developmental stages. One set of item parameters expresses relative difficulties among items within item classes, which, like Rasch item difficulty parameters, are the same for people in all stages. A second set of parameters quantifies patterns that the Rasch model cannot express: differences in relative difficulties between item classes for people in different stages, such as the difficulty reversals mentioned above. Saltus is effectively a mixture of standard Rasch models.

Mislevy and Verhelst (in press) have discussed mixture models more generally, listing assumptions, laying out general models, and suggesting estimation procedures. They emphasize situations in which different subjects follow different strategies, pointing out that instructional decisions can depend on how students solve problems, not just how many they solve. The salient features of items are those that can differentiate among users of different strategies, mental models, or conceptions about key relationships. An examinee is characterized by the probabilities that she employed the various alternative strategies, and a conditional estimate of proficiency under each. Measurement with such a model can indicate change that is either quantitative (e.g., the examinee employed Strategy A on both occasions, but more effectively at the second) or qualitative (e.g., she used Strategy A before instruction but Strategy B afterwards).

Latent class models. Although models with continuous latent variables have dominated educational measurement, Lazarsfeld (1950) introduced models with categorical latent variables nearly half a century ago. Most educational applications of latent class models have been in "mastery" testing; one attempts to infer an examinee's unobservable state--master or nonmaster--on the basis of observable responses (Macready and Dayton, 1977, 1980). In the more recent "binary skills" models (Haertel, 1984), examinees are classified in terms of which of a set of skills they possess. This "true" classification is unobservable. Items are classified

according to which of the skills they require for solution. This classification is known. Ideally, an examinee responds correctly to only and exactly those items that require skills he or she possesses. The stochastic parameters of the model reflect departures from this ideal.

Except in the special case of mastery testing, computational constraints have limited applications of latent class models to no more than about ten items until recently. Information about skill profiles in groups can be gleaned from such data, but individuals' skills could not be inferred accurately. Improved computational procedures have opened the door to applications with 50 or 60 items (e.g., Paulson, 1986; Yamamoto, 1987), and work with structurally similar models in expert systems holds promise of handling much larger problems (Lauritzen and Spiegelhalter, 1988). Progress in this direction is vital to educational applications, since these inferences demand more data than low-to-high proficiency inferences. Moreover, adaptive testing, which made IRT measurement more efficient, will be able to make latent class measurement practicable (Dayton and Macready, 1989; Falmagne and Doignon, 1988).

Componential models. The models described above were introduced with right/wrong test items, which, if constructed carefully, yield response patterns that differentiate examinees who tackle them in different ways. Richer information can be accumulated if it is possible to track intermediate products of solution. Consider, for example, a situation in which the binary skills model applies. Inferences about skill profiles can be stronger if one can be see which subtasks were attempted and their outcomes: overall correctness can result from one sequence of correct operations or another, or a fortuitous mixture of correct and incorrect operations; overall incorrectness can be caused by a poor plan of attack, or a flawed execution of a good plan. Early implementations of these ideas have been worked out by Embretson (1983, 1985b) and Samejima (1983).

All of the models discussed above--tectonic plate, latent class, and componential models--exhibit the same cardinal feature: they support inferences about proficiencies other than just low-to-high ability because, and only because, the user specifies theoretically salient patterns of response other than just less-to-more correct answers. Current implementations require expertise in statistics as well as in the substantive area. Test theory researchers must embed these approaches in generally applicable computer routines, or shells, so that a broader range of users can put them into practice in the substantive areas.[3]

Beyond Right/Wrong, Multiple-Choice Items

Currently IRT is used almost exclusively to draw inferences about a low-to-high proficiency variable from responses to multiple-choice test items. The preceding section discussed how, even with multiple-choice data, one can found inferences upon radically different conceptions of proficiency. Inferences can be made yet stronger, and decision-making more efficient, if different kinds of data can be collected.

We have mentioned the possibility of exploiting the identity of incorrect responses to multiple-choice items, for when particular misconceptions are probed in more than one item and we wish to infer how an examinee is approaching tasks. IRT models that distinguish among incorrect alternatives have been discussed by Bock (1972), Masters (1982), Samejima (1979), and Thissen and

---

[3] Similar diffusion processes have already occurred in two area related to test theory. The first is IRT itself. In the 1960's, only a handful of mathematically talented researchers could use IRT; now IRT is widely used by practitioners by virtue of production programs such as LOGIST (Wingersky, Barton, and Lord, 1982), BILOG (Mislevy and Bock, 1983), and BICAL (Wright, Mead, and Bell, 1980). The second area is that of linear structural relationships among variables with measurement error. Proposing such a model and solving the equations was once practically grounds for a Nobel prize in economics; now anyone with access to the LISREL computer program (Joreskog and Sorbom, 1986) can routinely carry out analyses undreamed of a few decades ago.

Steinberg (1984). These papers show how to connect observations more complex than right/wrong to the standard psychological model of low-to-high proficiency. The same machinery for the observational aspect of modeling can be used when the psychological aspect is an alternative cognitive model. Embretsen (1983, 1985b) and Masters (Masters and Mislevy, 1989) have taken some initial steps in this direction.

Because data collected on computers can provide response time routinely, response latency can also be exploited. Response latencies are particularly pertinent to inferences about automaticity; a correct answer arrived at through a laborious conscious process can have different instructional implications than the same response obtained through automatized processes. Response latencies can also be used in conjunction with correctness to design items that differentiate among examinees who use different strategies. Many quantitative items in the SAT, for example, can be solved either by a "brute force" calculation or by a simple calculation if a key relationship is recognized; "correct and fast" suggests the insightful solution. Scheiblechner (1985) and Thissen (1983) show how to use response times to measure low-to-high proficiency. Their methods of linking observed responses to expected responses could be applied with an alternative cognitive model for expected responses.

Beyond Tester-Controlled Observational Settings

Traditional educational tests present small, closed-form problems, isolated and packaged more neatly than the problems people encounter in life. Real-world tasks require one to recognize a problem space; to plan strategies, to take initial steps, and gather additional information; and, observing preliminary results, to determine which direction to proceed. Controlling the observational setting in testing to some degree is probably unavoidable in a decision-making system applied routinely to many learners. Controlled simulation tasks strike a compromise between the rigid, tester-controlled observational setting of

traditional tests and the wholely unstructured observation of performance in natural settings.

The most work in this area has been carried out in the arena of medical education in the form of "patient management problems," or PMPs (Assmann, Hixon, and Kacmarek, 1979). A simulated patient (through a written or oral dialogue, or as a live actor or a computer model) presents the examinee with initial symptoms; the examinee requests tests, considers their results, prescribes treatments, and monitors their effects, generally attempting to identify and treat the initially unknown disease. Despite their appeal as evocators of critical problem-solving skills, PMPs do not seem to provide reliable data from the perspective of standard test theoretic techniques (McGuire, 1985). For the same amount of testing time, reliability coefficients of PMP scores prove disappointingly low compared with multiple-choice tests.

A possible explanation of this result is that standard test theory analyses of PMP data are not looking for the right patterns. They look at simple additive combinations of single outcomes, rather than relationships that might suggest associations among facts in examinees' schema, or indicate the use of effective or ineffective problem-solving strategies. A distinct stream of medical research, however, does address these relationships: "expert systems" that help health care workers with diagnostic problems (e.g., Pope, 1981; Shortliffe et al., 1973).

An expert system representation of a health care area is build around associations among unobservable disease states, observable symptoms and test results, and outcomes of treatments. Some expert systems express these associations through "fuzzy logic" (Zadeh, 1983) or "belief functions" (Shafer, 1976), but the ones that use conditional probabilities (Spiegelhalter, 1986) are extensions of the latent class models discussed above. In an educational setting, associations would be delineated among substantive concepts, strategies, observable outcomes, and prescribed instruction (Clancey, 1988).

BEST COPY AVAILABLE

There are two levels at which expert systems could be implemented in educational settings. The first appears more amenable to end-of-course or macro-level decision-making, while the second seems better suited to an ongoing instructional system.

In the first, simpler, approach, an expert system is built only for a "correct" model. An examinee's responses are evaluated in terms of their efficacy at each decision point as compared with the best possible action given present information. If scores were also available from a standard multiple-choice test of knowledge, one could distinguish performance problems caused by strategic errors from those caused by knowledge deficiencies.

In the second, more ambitious, approach, not only would a correct expert system be built, but examinees' possibly "inexpert systems" would be inferred. Perhaps the best known example of this type is Anderson's (Anderson and Reiser, 1985) computer programming tutor. Although more individualized instructional prescriptions can be made in this way, inferring even selected aspects of examinees' schema and strategies requires far more data than does comparing performance to a fixed expert model. A successful system of this type would probably require a more constrained problem space and more extensive interactions of the learner with the simulation.

Conclusion

Einstein's theory of relativity revolutionized physics, but it extended rather than supplanted Newton's laws of motion. Classical mechanics still works just fine, thank you, for building bridges, planning billiards shots, and figuring out how to stand up from a overstuffed easy chair. And as long as educators are called upon to make the macro-level, linearly-ordered decisions that engendered standard test theory, standard test theory will continue to be useful, and will continue to be used. Recent developments in technology, however, provide opportunities for decision making at the micro-level more frequently and for larger numbers of students than ever before; recent developments in

education and psychology give us conceptions of competence and learning that can be used to guide these decisions.

Researchers in education and psychology have begun to lay the theoretical groundwork to link testing with the cognitive processes of learning. Meanwhile, researchers in measurement and statistics have made breakthroughs in inferential procedures for the models of standard test theory. To inform modern educational decisions requires drawing together the insights from these two strands of research--the twin foundations of a new test theory.

References

Anderson, J.R., and Reiser, B.J. (1985). The LISP tutor. Byte, 10, 159-175.

Assmann, D.C., Hixon, S.H., and Kacmarek, R.M. (1979). Clinical simulations for respiratory care workers. Chicago: Year Book Medical Publishers.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-52.

Bock, R.D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM-algorithm. Psychometrika, 46, 443-459.

Bock, R.D., Mislevy, R.J., and Woodsen, C.E.M. (1982). The next stage in educational assessment. Educational Researcher, 11, 4-11, 16.

Bloom, B.S. (1976). Human characteristics and school learning. New York: McGraw-Hill.

Braun, H.E. (1988). A new approach to avoiding problems of scale in interpreting trends in mental measurement data. Journal of Educational Measurement, 25, 171-191.

Caramazza, A., McCloskey, M., and Green, B. (1981). Naive beliefs in "sophisticated" subjects: Misconceptions about the trajectories of objects. Cognition, 9, 117-123.

Chi, M.T.H., Feltovich, P., and Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5, 121-152.

Choppin, B. (1976). Recent developments in item banking. In D.N. de Gruijter and L.J. van der Kamp (Eds.), Advances in psychological and educational measurement. London: Wiley.

Clancey, W.J. (1988). The role of qualitative models in instruction. In J. Self (Ed.), Artificial intelligence and

human learning: Intelligent computer-aided instruction. London: Chapman and Hall.

Cronbach, L.J., and Furby, L. (1970). How should we measure "change"--Or should we? Psychological Bulletin, 74, 68-80.

Cronbach, L.J., Gleser, G.C., Nanda, H., and Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Curtis, M.E., and Glaser, R. (1983). Reading theory and the assessment of reading achievement. Journal of Educational Measurement, 20, 133-147.

Dawes, R.M., and Corrigan, B. (1974). Linear models in decision making. Psychological Bulletin, 81, 95-106.

Dempster, A.P, Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39, 1-38.

Embretson, S.E. (1983). A general latent trait model for response processes. Psychometrika, 49, 175-186.

Embretson, S.E. (Ed.) (1985a). Test design: Developments in psychology and psychometrics. Orlando, FL: Academic Press.

Embretson, S.E. (1985b). Multicomponent latent trait models for test design. In S.E. Embretson (Ed.), Test design: Developments in psychology and psychometrics. Orlando, FL: Academic Press.

Falmagne, J.-C., and Doignon, J.-P. (1988). A class of stochastic procedures for the assessment of knowledge. British Journal of Mathematical and Statistical Psychology, 41, 1-23.

Ferguson, G.A. (1942). Item selection by the constant process. Psychometrika, 7, 19-29.

Fischer, G.H. (1983). Logistic latent trait models with linear constraints. Psychometrika, 48, 3-26.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. American Psychologist, 36, 923-936.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 118, 519-521.

Green, B. F. (1978) In defense of measurement. American Psychologist, 33, 664-670.

Greeno, J.G. (1976). Cognitive objectives of instruction: Theory of knowledge for solving problems and answering questions. In D. Klahr (Ed.), Cognition and instruction. Hillsdale, NJ: Wiley.

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

Guttman, L. (1944). A basis for scaling qualitative data. American Sociological Review, 9, 139-150.

Guttman, L. (1970). Integration of test design and analysis. Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service.

Haertel, E.H. (1984). An application of latent class models to assessment data. Applied Psychological Measurement, 8, 333-346.

Hunt, E., and MacLeod, C.M. (1978). The sentence-verification paradigm: A case study of two conflicting approaches to individual differences. Intelligence, 2, 129-144.

Kelley, T.L. (1923). Statistical methods. New York: Macmillan.

Kelley, T.L. (1927). Interpretation of Educational Measurements. New York: World Book.

Kuder, G.F., and Richardson, M.W. (1937). The theory of the estimation of test reliability. Psychometrika, 2, 151-160.

Lauritzen, S.L., & Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). Journal of the Royal Statistical Society, Series B, 50, 157-224.

Lawley, D.N. (1943). On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, Section A, 61, 273-287.

Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis. In S.A. Stouffer, L. Guttman,

E.A. Suchman, P.F. Lazarsfeld,, S.A. Star, and J.A. Clausen, Studies in social psychology in World War II, Volume 4: Measurement and Prediction. Princeton, NJ: Princeton University Press.

Lesgold, A.M., and Perfetti, C.A. (1978). Interactive processes in reading comprehension. Discourse Processes, 1, 323-336.

Lord, F.M., (1952). A theory of test scores. Psychometrika Monograph No. 7, 17 (4, Pt. 2).

Lord, F.M. (1958). Further problems in the measurement of growth. Educational and Psychological Measurement, 18, 437-454.

Lord, F.M. (1959). Statistical inference about true scores. Psychometrika, 24, 1-18.

Lord, F.M., (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F.M., and Novick, M.R. (1968). Statistical theories of mental test scores. Reading, Mass: Addison-Wesley.

Macready, G.B., and Dayton, C.M. (1977). The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 2, 99-120.

Macready, G.B., and Dayton, C.M. (1980). The nature and use of state mastery models. Applied Psychological Measurement, 4, 493-516.

Macready, G.B., and Dayton, C.M. (1989, March). The application of latent class models in adaptive testing. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Marshall, S.P. (1985). Using schema knowledge to solve story problems. Paper presented at the Office of Naval Research Contractors' Conference, San Diego, CA, December 1985.

Masters, G.N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

Masters, G.N., and Mislevy, R.J. (1989). New views of student learning: Implications for educational measurement. In N. Frederiksen, R.J. Mislevy, and I. Bejar (Eds.), Test theory for a new generation of tests. Hillsdale, NJ: Erlbaum.

McGuire, C.H. (1985). Medical problem-solving: A critique of the literature. Journal of Medical Education, 60, 587-595.

Mead, R.J. (1976). Analysis of fit to the Rasch model. Unpublished doctoral dissertation. University of Chicago.

Messick, S. (1984). The psychology of educational measurement. Journal of Educational Measurement, 23, 147-156.

Messick, S., Beaton, A.E., and Lord, F.M. (1983). National Assessment of Educational Progress reconsidered: A new design for a new era. NAEP Report 83-1. Princeton, NJ: National Assessment for Educational Progress.

Mislevy, R.J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. Applied Psychological Measurement, 11, 81-91.

Mislevy, R.J. (1988a). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. Applied Psychological Measurement, 12, 281-296.

Mislevy, R.J. (1988b). Randomization-based inferences about latent variables from complex samples. ETS Research Report RR-88-54-ONR. Princeton: Educational Testing Service.

Mislevy, R.J., and Sheehan, K.M. (in press). The role of collateral information about examinees in the estimation of item parameters. Psychometrika.

Mislevy, R.J., and Bock, R.D. (1983). BILOG: Item analysis and test scoring with binary logistic models [computer program]. Mooresville, IN: Scientific Software, Inc.

Mislevy, R.J., and Verhelst, N. (in press). Modeling item responses when different subjects employ different solution strategies. Psychometrika.

Mislevy, R.J., and Wu, P-K. (1988). Inferring examinee ability when some item responses are missing. ETS Research Report RR-88-48-ONR. Princeton: Educational Testing Service.

Novick, M.R. (1966). The axioms and principle results of classical test theory. Journal of Mathematical Psychology, 3, 1-18.

Novick, M.R., and Jackson, P.H. (1974). Statistical methods for educational and psychological research. New York: McGraw-Hill.

Paulson, J.A. (1986). Latent class representation of systematic patterns in test responses. Technical Report ONR-1. Portland, OR: Psychology Department, Portland State University.

Pope, H.E. (1981). Heuristic methods for imposing structure on ill-structured problems: The structuring of medical diagnostics. In P. Szolovitz (Ed.), Artificial intelligence in medicine. Boulder: Westview Press.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research.

Samejima, F. (1979). A new family of models for the multiple-choice item. ONR Research Report 79-4. Knoxville, TN: University of Tennessee.

Samejima, F. (1983). A latent trait model for differential strategies in cognitive processes. ONR Research Report 83-1. Knoxville, TN: University of Tennessee.

Scheiblechner, H. (1972). Das lernen und losen komplexer denkaufgaben. Zeitschrift fur experimentalle und Angewandte Psychologie, 19, 476-506.

Scheiblechner, H. (1985). Psychometric models for speed-test construction: The linear exponential model. In S.E. Embretson (Ed.), Test design: Developments in psychology and psychometrics. Orlando, FL: Academic Press.

Siegler, R.S. (1981). Developmental sequences within and between concepts. Monograph of the Society for Research in child Development, Serial No. 189, 46.

Shafer, G. (1976). A mathematical theory of evidence. Princeton: Princeton University Press.

Shortliffe, E.H., Axline, S.G., Buchanan, B.G., Merigan, T.C., and Cohen, S.W. (1973). An artificial intelligence program to

advise physicians regarding antimicrobial therapy. Computers in Biomedical Research, 6, 544-560.

Smith, R. (1986). Person fit in the Rasch model. Educational and Psychological Measurement, 46, 359- 372.

Spearman, C. (1904a). The proof and measurement of association between two things. American Journal of Psychology, 15, 72-101.

Spearman, C. (1904b). "General intelligence" objectively determined and measured. American Journal of Psychology, 15, 201-292.

Spearman, C. (1907). Demonstration of formulae for true measure of correlation. American Journal of Psychology, 18, 161-169.

Spearman, C. (1910). Correlation calculated with faulty data. British Journal of Psychology, 3, 271-295.

Spearman, C. (1913). Correlations of sums and differences. British Journal of Psychology, 5, 417-426.

Spiegelhalter, D.J. (1986). Probabilistic reasoning in predictive expert systems. In L.W. Kanal and J. Lemmer (Eds.), Artificial intelligence and statistics. Amsterdam: North-Holland.

Tatsuoka, K.K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 20, 345- 354.

Theunissen, T.J.J.M. (1985). Binary programming and test design. Psychometrika, 50, 411-420.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D.J. Weiss (Ed.), New horizons in testing: Latent trait test theory and computerized adaptive testing. New York: Academic Press.

Thissen, D., and Steinberg, L. (1984). A response model for multiple choice items. Psychometrika, 47, 201-214.

Thorndike, E.L., Bregman, E.O., Cobb, M.V., and Woodyard, E. (1926). The measurement of intelligence. New York: Columbia Teachers College, Bureau of Publications.

Thurstone, L.L. (1925). A method of scaling psychological and educational tests. Journal of Educational Psychology, 16, 433-451.

Thurstone, L.L. (1928). The measurement of opinion. Journal of Abnormal and Social Psychology, 22, 415-430.

Tsutakawa, R.K. (1984). Estimation of two-parameter logistic item response curves. Journal of Educational Statistics, 9, 263-276.

Tsutakawa, R.K., and Soltys, M.J. (1988). Approximation for Bayesian ability estimation. Journal of Educational Statistics, 13, 117-130.

Tucker, L. R (1946). Maximum validity of a test with equivalent items. Psychometrika, 11, 1-13.

Vosniadou, S., and Brewer, W.F. (1987). Theories of knowledge restructuring in development. Review of Educational Research, 57, 51-67.

Weiss, D.J. (1984). Application of computerized adaptive testing to educational problems. Journal of Educational Measurement, 21, 361-376.

Wilson, M.R. (1985). Measuring stages of growth: A psychometric model of hierarchical development. Occasional Paper No. 19. Hawthorne, Australia: Australian Council for Educational Research.

Wilson, M.R. (1989). Saltus: A psychometric model of discontinuity in cognitive development. Psychological Bulletin, 105, 276-289.

Wingersky, M.S., Barton, M.A., and Lord, F.M. (1982). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

Wright, B.D. (1968). Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service.

Wright, B.D., Mead, R.J., and Bell, S.R. (1980). BICAL: Calibrating items with the Rasch model. Research Memorandum

23C. Chicago: Statistical Laboratory, Department of Education, University of Chicago.

Yamamoto, K. (1987). A model that combines IRT and latent class models. Unpublished doctoral dissertation, University of Illinois, Champaign-Urbana.

Zadeh, L.A. (1983). The role of fuzzy logic in the management of uncertainty in expert systems. Fuzzy Sets and Systems, 11, 199-227.